ÁP DỤNG RỪNG NGẪU NHIÊN TRONG HỌC MÁY DỰ ĐOÁN CHẤT LƯỢNG RƯỢU VANG

APPLICATION OF RANDOM FOREST IN MACHINE LEARNING TO PREDICT WINE QUALITY

Đỗ Thị Kim Dung[1,*], Lê Đình Phú Cường[2], Viên Thanh Nhã[3], Lê Đình Hồng Mạnh[4], Phạm Văn Cường[4], Phan Đức Thiện[5], Phạm Thành Công[4], Lê Việt Anh[4]

DOI: https://doi.org/10.57001/huih5804.2023.107

TÓM TẮT

Hiện nay, học máy được ứng dụng ngày càng nhiều vào đời sống. Máy móc cũng có thể hỗ trợ con người lựa chọn các sản phẩm phù hợp. Nhà sản xuất thì muốn sản xuất một mẫu rượu phù hợp cho người tiêu dùng và khách hàng thì muốn có một mẫu rượu phù hợp với lựa chọn của mình. Hơn nữa, chất lượng của rượu vang không chỉ phụ thuộc vào một yếu tố nhất định mà nó phụ thuộc vào nhiều yếu tố. Nếu dựa vào cách thủ công để dự đoán chất lượng thì mất rất nhiều thời gian. Dựa vào nhu cầu thực tế đó trong nghiên cứu này chúng tôi đề xuất sử dụng 3 phương pháp DT (Decision Tree), SVM (Support Vector Machine), RF (Random Forest) trong học máy để dự đoán rượu vang. Dữ liệu rượu vang được sử dụng làm cơ sở đánh giá có 1599 dòng, mỗi dòng có 12 cột. Kết quả thực nghiệm cho thấy phương pháp RF cho kết quả tốt nhất, dựa vào kết quả này chúng tôi xây dựng trang web dự đoán chất lượng rượu.

Từ khóa: *Học máy RF, chất lượng, dự đoán.*

ABSTRACT

Currently, machine learning is applied more and more in life. Machines can also assist humans in choosing the right products. The producer wants to produce a suitable wine for the consumer and the customer wants a suitable wine of his choice. More than half, the quality of wine depends not only on a certain factor, but it depends on many factors. If you rely on manual methods to predict the quality, it takes a lot of time. Based on that actual need in this study, we propose to use 3 methods DT (Decision Tree), SVM (Support Vector Machine), RF (Random Forest) in machine learning to predict wine. The wine data used as the basis of the assessment has 1599 lines, each with 12 columns. Experimental results show that RF method gives the best result, based on this result we build a wine quality prediction website.

Keywords: *Machine learning RF, quality, prediction.*

[1]Khoa Công nghệ thông tin, Trường Đại học Phan Thiết
[2]Khoa Công nghệ thông tin, Trường Đại học Yersin Đà Lạt
[3]Khoa Công nghệ thông tin, Trường Đại học Thủy Lợi - Phân hiệu miền Nam
[4]Trường Đại học Công nghiệp Hà Nội
[5]Trường Đại học Sư Phạm Kỹ thuật Nam Định
[*]Email: dtkdung@upt.edu.vn
Ngày nhận bài: 20/02/2023
Ngày nhận bài sửa sau phản biện: 29/3/2023
Ngày chấp nhận đăng: 15/6/2023

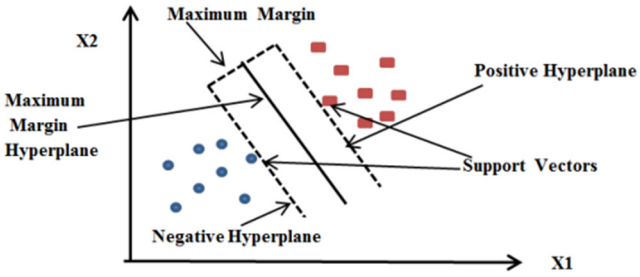
1. GIỚI THIỆU

Chất lượng của rượu vang được người tiêu dùng và nhà sản xuất rất quan tâm. Có rất nhiều loại rượu vang trên thị trường, nó đa dạng về màu sắc, hình dạng và nhiều đặc tính khác nhau. Trong đó có một vài đặc tính sẽ thay đổi theo thời gian và có thể làm kém đi chất lượng của rượu. Bên cạnh đó nếu tiêu thụ nhiều sản phẩm kém chất lượng sẽ làm ảnh hưởng nghiêm trọng đến sức khỏe người sử dụng. Một ứng dụng tự động dự đoán chất lượng rượu là rất cần thiết.

Có nhiều cách giải quyết vấn đề này, tuy nhiên đa phần là giải quyết thủ công còn phụ thuộc vào yếu tố con người là chính. Trong những năm gần đây một số nhà nghiên cứu đã sử dụng học máy để giải quyết vấn đề này, chẳng hạn như: Terence đã sử dụng bộ dữ liệu chất lượng rượu vang đỏ tại trang kaggle.com và dùng nhiều mô hình phân loại để dự đoán xem một mẫu rượu vang đỏ có tốt hay không. Devika để dự đoán chất lượng rượu vang, các nhà nghiên cứu đã sử dụng Logistic Regression, Stochastic Gradient Descent, Support Vector Classifier và Random Forest [2]. Phân tích chứng minh rằng chất lượng được cải thiện khi lượng đường dư ở mức vừa phải và không thay đổi đột ngột, cho thấy đặc điểm này không quan trọng bằng các đặc điểm khác như rượu và axit xitric. Bài toán dự đoán chất lượng rượu bằng so sánh các phương pháp học máy và đưa ra dự đoán cao được chúng tôi áp dụng trong bài báo này.

2. PHƯƠNG PHÁP

2.1. Học máy véc tơ hỗ trợ - Support Vector Machine (SVM)



Hình 1. Bộ phân loại SVM

SVM là một kỹ thuật học máy có giám sát có thể sử dụng để giải quyết các vấn đề trong phân loại và hồi quy. SVM tạo ra một siêu phẳng để phân tách các lớp trong không gian n chiều, một siêu phẳng có thể là ranh giới quyết định hoặc một số đường. Vectơ hỗ trợ là các điểm dữ liệu hoặc vectơ gần siêu phẳng nhất và ảnh hưởng đến vị trí của siêu phẳng.

Margin của 1 lớp là khoảng cách từ các điểm gần nhất của lớp đó tới mặt phân chia. Margin của 2 lớp phải bằng nhau và phải lớp nhất có thể. Cách tính lại Margin sẽ là:

$$\text{Margin} = \min \frac{y_n(w^T x_n + b)}{\|w\|_2} \quad (1)$$

Với $y_n = \pm 1$ là nhãn của điểm dữ liệu

Hiệu ứng phân lớp tốt hơn đối với Margin rộng vì 2 lớp được phân chia cụ thể. Tìm đường phân chia là lớn nhất đối với Margin giữa 2 lớp.

$$(w, b) = \arg\max_{w,b} \left\{ \min_n \frac{y_n(w^T x_n + b)}{\|w\|_2} \right\}$$

$$= \arg\max_{w,b} \left\{ \frac{1}{\|w\|_2} \min_n y_n (w^T x_n + b) \right\} \quad (2)$$

Với mọi n ta có: $y_n(w^T x_n + b) \geq 1$

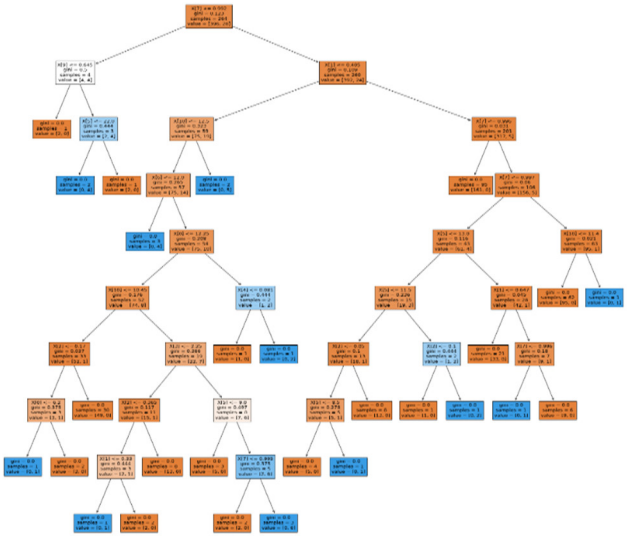
Vậy tối ưu có ràng buộc cho bài toán sẽ là:

$$(w, b) = \arg\max_{w,b} \frac{1}{\|w\|_2} \quad (3)$$

Thỏa mãn: $y_n(w^T x_n + b) \geq 1, \forall n = 1,2, \dots, N$

2.2. Cây quyết định - Decision Tree (DT)

Cây quyết định là một mô hình học có giám sát, các bài toán phân loại dữ liệu và hồi quy đều có thể áp dụng được. Chuẩn hóa dữ liệu và không bắt buộc chia tỷ lệ khi sử dụng cây quyết định. Một thay đổi nhỏ trong dữ liệu có thể dẫn đến sự thay đổi đáng kể trong cấu trúc của cây quyết định tối ưu.



Hình 2. Mô hình cây quyết định phân loại rượu vang

Trong hình 2 mô tả cây quyết định với các thuộc tính trong tập dữ liệu được đại diện cho các nút bên trong và các quy tắc quyết định được đại diện cho các nhánh.

2.3. Rừng ngẫu nhiên - Random Forest (RF)

RF là một phương pháp học tập sử dụng việc xây dựng nhiều cây quyết định. Dựa trên phần lớn các cây để thực hiện lựa chọn khu rừng ngẫu nhiên. Bằng cách chọn ngẫu nhiên các dòng từ một tập dữ liệu để RF tạo ra n số cây quyết định. RF thu thập các dự báo từ mỗi cây và dự đoán kết quả cuối cùng thay vì dựa vào một cây quyết định duy nhất. Độ chính càng cao thì đòi hỏi số lượng cây trong rừng càng lớn. Cách tạo mô hình RF, cần xác định xây dựng được bao nhiêu cây. Bootstrap là từ các dữ liệu ngẫu nhiên mỗi cây được xây dựng. Trong thống kê và học máy mẫu bootstrap được sử dụng phổ biến.

Hình 3. Mô hình rừng ngẫu nhiên

3. XÂY DỰNG, HUẤN LUYỆN MÔ HÌNH

3.1. Thu thập, tiền xử lý dữ liệu

Thực hiện việc xây dựng và huấn luyện mô hình, chúng tôi sử dụng bộ dữ liệu của tác giả Cortez (2009) tại Kaggle. Kho dữ liệu rượu vang gồm 1599 dòng với 12 cột được thể hiện bảng 1.

Bảng 1. Thông tin bộ dữ liệu rượu

TT	Thuộc tính	Giải thích
1	Fixed acidity	Độ axit cố định (g/dm^3)
2	Volatile acidity	Độ bay hơi axit (g/dm^3)
3	Axit citric	Axit citric (g/dm^3)
4	Residual sugar	Đường dư (g/dm^3)
5	Chlorides	Clorua
6	Free sulfur dioxide	Lưu huỳnh đioxit tự do (mg/dm^3)
7	Total sulfur dioxide	Tổng lượng lưu huỳnh điôxít (mg/dm^3)
8	Density	Mật độ (g/cm^3)
9	pH	Độ pH
10	Sulphates	Muối Sulfat (g/dm^3)
11	Alcohol	Cồn (%vol)
12	Quality	Chất lượng

Đánh giá rượu đó tốt hay không tốt (1 hay 0). Rượu có điểm từ 6 trở lên là rượu tốt, còn lại là không tốt. Bên cạnh đó dữ liệu chọn để huấn luyện là 80% và kiểm thử là 20% được lấy một cách ngẫu nhiên.

3.2. Xây dựng và huấn luyện mô hình

Sử dụng ba phương pháp học máy SVM, DT và RF để xây dựng và huấn luyện mô hình dự đoán chất lượng rượu dựa trên 11 thuộc tính được trình bày trong bảng 1. Tổng thể mô hình dự đoán được thể hiện chi tiết ở hình 4.

4. KẾT QUẢ THỰC NGHIỆM

Quá trình thực nghiệm sử dụng ngôn ngữ lập trình python, các gói python được nhập để hỗ trợ trong học máy là seaborn, tensorflow, matplotlib, pandas, numpy, gradio… Sau khi tiền xử lý dữ liệu bộ dữ liệu chuyển thành vector và huấn luyện bằng ba phương pháp học máy là SVM, RF và DT. Có tổng cộng 12 biến được sử dụng. Trong đó, biến chất lượng được coi là biến phụ thuộc và 11 biến khác được giả định là yếu tố dự báo. Đánh giá kết quả trên 4 độ đo: độ phân loại chính xác, độ chính xác, độ bao phủ và độ đo F1 score. Công thức tính các độ đo này:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (4)$$

Accuracy: Nó xác định tần suất chính xác mô hình dự đoán đầu ra. Đối với bài toán phân loại thông số này rất quan trọng.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (5)$$

Precision là khả năng của bài toán phân loại mà giá trị negative không được gán cho mẫu positive. Đối với mỗi class được tính theo công thức (5).

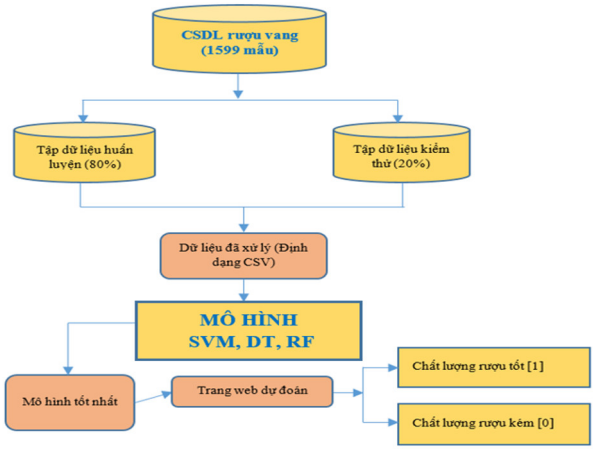
$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (6)$$

Recall là khả năng một bài toán phân lớp các mẫu được tìm ra. Đối với mỗi class nó được tính theo công thức (6)

F1 score là chỉ số trung hòa giữa giá trị Precision và Recall.

$$\text{F1} - \text{score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision+Recall}} \quad (7)$$

Trong đó: True negative (TN), True positive (TP), False negative (FN), False positive (FP)



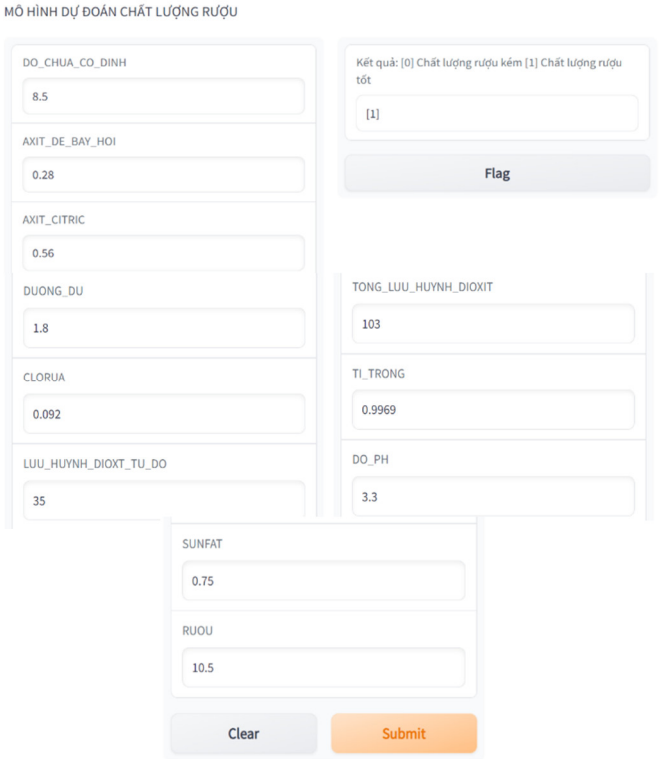
Hình 4. Tổng thể mô hình dự đoán chất lượng rượu

Bảng 2. Kết quả của 3 phương pháp DT, SVM, RF

	DT			SVM			RF		
	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
0	0,75	0,72	0,73	0,75	0,73	0,74	0,83	0,76	0,79
1	0,72	0,75	0,73	0,72	0,74	0,73	0,77	0,74	0,80
accuracy			0,73			0,73			0,80
macro avg	0,73	0,73	0,73	0,74	0,74	0,73	0,80	0,80	0,80
weighted avg	0,73	0,73	0,73	0,74	0,73	0,74	0,80	0,80	0,80

Kết quả tỷ lệ chính xác của ba phương pháp DT, SVM, RF lần lượt là 0,73; 0,73; 0,80 được thể hiện đầy đủ trong bảng 2.

Theo bảng 2 thì RF có độ chính xác đạt (80%), hai phương pháp còn lại đạt (73%) độ chính xác. Như vậy trong ba phương pháp thì RF có độ chính xác cao nhất. Tiếp tục xây dựng trang web cho phép người dùng nhập vào 11 thông số: Độ chua cố định, Axit dễ bay hơi, Axit citric, Đường dư, Clorua, Lưu huỳnh dioxit tự do, Tổng lưu huỳnh dioxit, Tỷ trọng, Độ pH, Sunfat, Cồn của một mẫu bất kỳ. Hệ thống sẽ bắt đầu tiền xử lý thông tin rồi chuyển qua mô hình RF sau đó hiển thị kết quả dự đoán. Nếu kết quả là [0] thì chất lượng rượu kém và ngược lại [1] là chất lượng rượu tốt. Kết quả được minh họa trong hình 5.



Hình 5. Trang web dự đoán chất lượng rượu vang

5. KẾT LUẬN

Trong bài báo này, nhóm tác giả sử dụng ba phương pháp SV, RF, SVM để dự đoán chất lượng rượu vang bằng

mô hình rừng ngẫu nhiên. Bên cạnh đó số liệu để mô phỏng được tổng hợp từ kết quả nghiên cứu đã được công bố trên các tạp chí uy tín. Trong ba phương pháp thì RF cho kết quả tốt nhất để dự đoán. Từ mô hình RF này, tiếp tục xây dựng trang Web dự đoán chất lượng rượu trực tuyến. Nó giúp ích cho người tiêu dùng, giảm thiểu số vụ gian lận trong ngành rượu và giúp các công ty giảm các sai sót so với dự đoán thủ công. Trong tương lai, một tập dữ liệu khổng lồ có thể được sử dụng để nghiên cứu và có nhiều phương pháp học máy khác nhau để dự đoán chất lượng rượu cho độ chính xác cao hơn.

TÀI LIỆU THAM KHẢO

[1]. Binh T. H., 2015. *Ung dung Random Forest de tu van chon lo trinh hoc trong hoc che tin chi*. Master Thesis, The University of Da Nang.

[2]. Devika P., Aakanksha M., Sachin B., 2019. *Wine Quality Prediction using Machine Learning Algorithms*. International Journal of Computer Applications Technology and Research, Volume 8, Issue 09, 385-388, ISSN: 2319–8656.

[3]. Zeng Y., Liu Y., Wu L., Dong H., Zhang Y., Guo H., Guo Z., Wang S., Lan Y., 2018. *Evaluation and Analysis Model of Wine Quality Based on Mathematical Model*. Studies in Engineering and Technology, 6(1), 6, doi:10.11114/set.v6i1.3626.

[4]. Er Y., 2016. *The Classification of White Wine and Red Wine According to Their Physicochemical Qualities*. International Journal of Intelligent Systems and Applications in Engineering, 4 (1), 23–26.

[5]. Tom M. Mitchell, 1997. *Machine Learning*. McGraw Hill.

[6]. https://towardsdatascience.com/predicting-wine-qualitywith-several-classification-techniques-179038ea6434.

[7]. https://www.kaggle.com/code/ashishkumarbehera/red-wine-quality-classificaton

[8]. https://gradio.app/

[9]. https://dev.to/leading-edje/machine-learning-and-wine-quality-finding-a-good-wine-using-multiple-classifications-4kho

[10]. https://machinelearningcoban.com/tabml_book/ch_model/random_forest.html

AUTHORS INFORMATION

Do Thi Kim Dung[1], Le Dinh Phu Cuong[2], Vien Thanh Nha[3], Le Dinh Hong Manh[4], Pham Van Cuong[4], Phan Duc Thien[5], Pham Thanh Cong[4], Le Viet Anh[4]

[1]Faculty of Information Technology, University of Phan Thiet, Vietnam

[2]Faculty of Information Technology, Yersin University, Vietnam

[3]Faculty of Information Technology, ThuyLoi University - Southern Campus, Vietnam

[4]Hanoi University of Industry, Vietnam

[5]Namdinh University of Technology and Education, Vietnam